Vol. 21 no. 7 2005, pages 846–852
doi:10.1093/bioinformatics/bti072

Genome analysis

ColorHOR—novel graphical algorithm for fast scan of alpha satellite higher-order repeats and HOR annotation for GenBank sequence of human genome

Vladimir Paar[1,*], Nenad Pavin[1], Marija Rosandić[2], Matko Glunčić[1], Ivan Basar[1], Robert Pezer[1] and Sonja Durajlija Žinić[3]

[1]Faculty of Science, University of Zagreb, Bijenička 32, 10000 Zagreb, Croatia, [2]Department of Internal Medicine, University Hospital Rebro, Kišpatićeva 12, Zagreb, Croatia and [3]Ruđer Bošković Institute, Bijenička 54, Zagreb, Croatia

Received on July 12, 2004; revised and accepted on September 15, 2004
Advance Access publication October 27, 2004

ABSTRACT

Motivation: GenBank data are at present lacking alpha satellite higher-order repeat (HOR) annotation. Furthermore, exact HOR consensus lengths have not been reported so far. Given the fast growth of sequence databases in the centromeric region, it is of increasing interest to have efficient tools for computational identification and analysis of HORs from known sequences.

Results: We develop a graphical user interface method, ColorHOR, for fast computational identification of HORs in a given genomic sequence, without requiring a priori information on the composition of the genomic sequence. ColorHOR is based on an extension of the key-string algorithm and provides a color representation of the order and orientation of HORs. For the key string, we use a robust 6 bp string from a consensus alpha satellite and its representative nature is tested. ColorHOR algorithm provides a direct visual identification of HORs (direct and/or reverse complement). In more detail, we first illustrate the ColorHOR results for human chromosome 1. Using ColorHOR we determine for the first time the HOR annotation of the GenBank sequence of the whole human genome. In addition to some HORs, corresponding to those determined previously biochemically, we find new HORs in chromosomes 4, 8, 9, 10, 11 and 19. For the first time, we determine exact consensus lengths of HORs in 10 chromosomes. We propose that the HOR assignment obtained by using ColorHOR be included into the GenBank database.

Availability: The program with graphical user interface application for ColorHOR is freely available at http://www.hazu.hr/KSA/colorHOR.html. It can be run on any platform on which wxPython is supported.

Contact: paar@hazu.hr

Supplementary information: http://www.hazu.hr/KSA/colorHOR.html.

INTRODUCTION

Alpha satellite DNA defined by a diverged ∼171 bp motif repeated in a head-to-tail tandem, was identified at centromeres of all primate chromosomes (Warburton and Willard, 1996). Human and other primate chromosomes contain alpha satellites hierarchically organized into higher-order repeat (HOR) or alphoid arrays, which were systematically studied by restriction enzyme analysis (Maio, 1971; Manuelidis and Wu, 1978; Willard, 1985; Waye and Willard, 1985; Willard and Waye, 1987a; Wevrick *et al*., 1992; Lee *et al*., 1997). Stretches of alpha satellites without additional sequence structure are known as monomeric (Wevrick *et al*., 1992; Alexandrov *et al*., 1993; Schueler *et al*., 2001).

Independent, ∼171 bp monomers of alpha satellite DNA generally exhibit substantial intermonomeric sequence divergence (20–40%), while HORs exhibit mutual sequence divergence of <5% (Maio, 1971; Manuelidis and Wu, 1978; Willard, 1985; Waye and Willard, 1985; Warburton and Willard, 1996; Puente *et al*., 1998; Wevrick *et al*., 1992; Lee *et al*., 1997).

Given the fast growth of sequence databases, it is of increasing interest to have tools for computational analysis of tandem repeats in a given sequence. Various algorithms for searching tandem repeats have been developed (Milosavljevic and Jurka, 1993; Benson and Waterman, 1994; Benson, 1999; Blanchard *et al*., 2000; Baldi and Baisnee, 2000; Borštnik *et al*., 1994; Castello *et al*., 2002; Hauth, 2002; Hauth and Joseph, 2002; Rosandić *et al*., 2003). However, despite significant progress in computational genomics, identification of complex repetitive patterns within anonymous DNA sequences in the presence of sizeable deviations from periodicity remains a challenge. In particular, for HORs the difficulties are largely due to imperfect patterns containing substitutions, insertions and deletions.

Recently, we introduced a novel key-string algorithm (KSA) for computational identification and study of HORs in clones containing alpha satellites (Rosandić *et al*., 2003). Key-string was used in KSA to segment computationally a given genomic sequence into key-string subsequences, resulting in an array of key-string lengths that enables identification of HORs. While a wide class of 3–6 bp key strings provide HORs, only for some key-strings, referred to as dominant, they are segmented simultaneously into both HORs and alpha monomers (Rosandić *et al*., 2003).

In this paper the KSA method is extended by developing a novel graphical method, the ColorHOR algorithm, providing a visual display. We note that coloring is widely used in some contexts of bioinformatics; for example, the subtle conservation patterns can be revealed by coloring according to the conservation of amino acid groupings (Taylor, 1986) or for consensus-based coloring of multiple

[*]To whom correspondence should be addressed.

alignments (Goodstadt and Ponting, 2001). A detailed presentation of ColorHOR results is given for chromosome 1 as an illustration. Applying ColorHOR we identify HORs in GenBank data of the whole human genome.

SYSTEM AND METHODS

ColorHOR algorithm

The graphical algorithm introduced and developed in this paper, based on the use of key-string, models a given genomic sequence onto a graph in such a way that an elegant graph-layout is obtained. This algorithm is very effective for long sequences, as for example for a whole chromosome or genome, and proceeds as follows.

Choice of key-string for human genome The only input to ColorHOR is a key-string which segments a given sequence both into HORs and alpha satellites. Here, we provide a simple prescription for the choice of key-string for human genome, using the most robust 6 bp strings from known alpha satellite consensus sequence. In fact, the most robust 6 bp string in human alpha satellite consensus sequence from Choo *et al*. (1991) is GAAACA; the corresponding average percentage of consensus bases in alpha satellite monomers of wide-ranging chromosomal origin is 92%. Only slightly less robust strings from alpha satellite consensus sequence from Choo *et al*. (1991) are AGAAAC, GAGCAG, AAACAC and AGAGAA. Similar robustness can be seen in more recent alpha satellite consensus (Romanova *et al*., 1996), and is also consistent with early consensus from Waye and Willard (1987). Finally, after completing our ColorHOR treatment of human genome, we checked that the same robustness is persistent in samples of tens of thousands of alpha satellite monomers from GenBank data.

With respect to the choice of 6 bp key string, we note that the number of possible variations for 6 bp strings is $4^6 = 4096$, leading, in general, to a spectrum of distance frequencies with pronounced peaks in a domain of HOR lengths up to a few thousands.

For the application of the ColorHOR algorithm to GenBank sequence of human genome, we choose the key-string TGTTTC, a reverse complement of GAAACA [for alpha satellite consensus sequence from Choo *et al*. (1991) the + strand was used].

By repeating the ColorHOR analysis of the GenBank sequence of the human genome using some other robust strings from alpha satellite consensus sequence as a key-string, we obtain results similar to those corresponding to the initial key string TGTTTC.

Computational construction of length–frequency distribution (N versus Δ diagram) Using the key-string, the GenBank genomic sequence is computationally segmented into key-string subsequences, each starting with the key-string. Distribution of frequency N versus key-string subsequence length Δ is displayed diagrammatically (N versus Δ diagram) (Fig. 1).

Computational construction of alpha staircase (N_c versus n diagram) for a whole chromosome In the next step, the cumulative frequency N_c of the key-string subsequence length $\Delta = 171$ bp, up to a position n, is computed and displayed diagrammatically along the chromosome sequence (N_c versus n diagram) (Fig. 2). Any local clustering of the 171 bp subsequence lengths along the sequence results in a steep increase (stair) in the N_c versus n graph. Such a graph along a whole chromosome sequence will be referred to as the alpha staircase. The location of each alpha satellite-containing clone in a chromosome sequence will be associated with a stair in the alpha staircase. This provides fast identification of clones and/or contigs containing alpha satellites.

Computational construction of colored bands and color-motif For each alpha satellite-containing clone, identified in the previous step, the corresponding length–frequency distribution (N versus Δ diagram) is displayed for the same key-string as used in the previous step for computation of the N versus Δ diagram of the whole chromosome sequence (Fig. 3). On the basis of the highest peaks we define the coloring rule: to each length corresponding to pronounced peaks a particular color is assigned. According to this coloring rule, the stripes displaying the corresponding key-string subsequences along the band of genomic sequence are colored. In this way, a colored band with repetitive color-motif is obtained at the location of each HOR-containing clone (Fig. 4).

A direct visual inspection easily reveals a repetitive color-motif, corresponding to HOR. Coordinates of HOR regions and of individual HOR copies, as well as larger deviations from consensus (insertions, deletions), are easily identified from the pattern of colored bands.

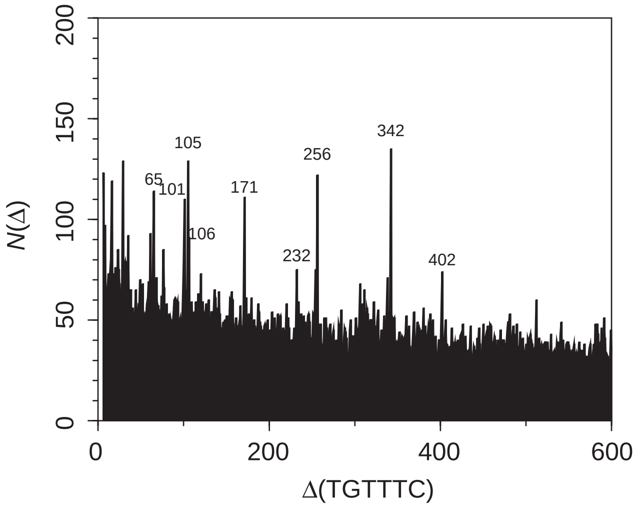

Fig. 1. Distribution of frequency N of subsequence lengths Δ for the GenBank sequence of chromosome 1. Key-string: TGTTTC. The unit of frequency (vertical axis) is an absolute number of subsequences of a given length that begin with the key-string, found by sequential search from one end of the chromosome to the other.

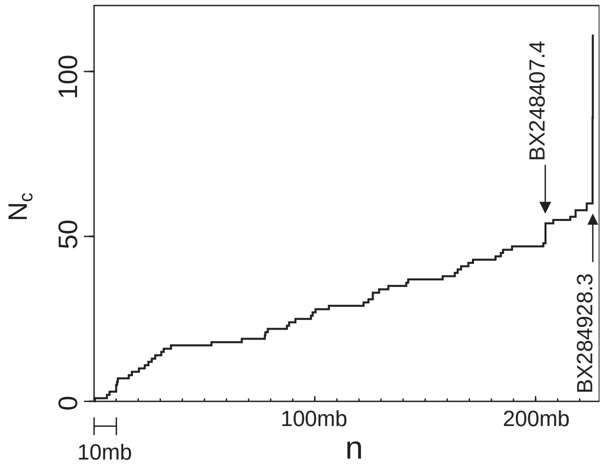

Fig. 2. Alpha staircase (cumulative frequency N_c versus position n of nucleotides along the GenBank sequence) for chromosome 1. Stairs in the alpha staircase correspond to locations of alpha satellite regions. The largest steps are at locations of BX284928.3 and BX248407.4. These BACs are not part of assembled chromosome 1 contigs; they are placed at the far right-hand side of the GenBank sequence.

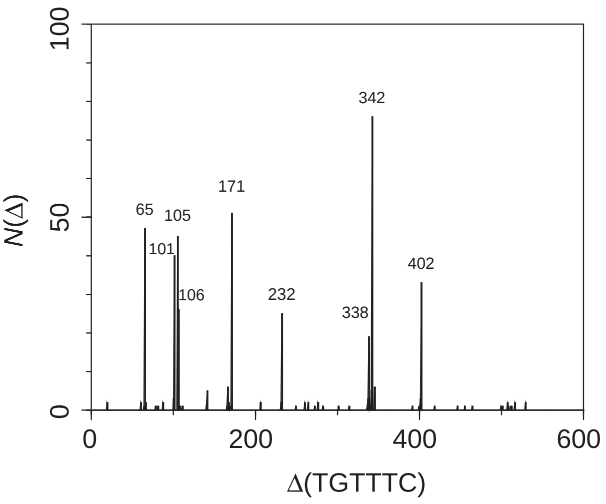

Fig. 3. Distribution of frequency N of subsequence lengths Δ for BX284928.3.

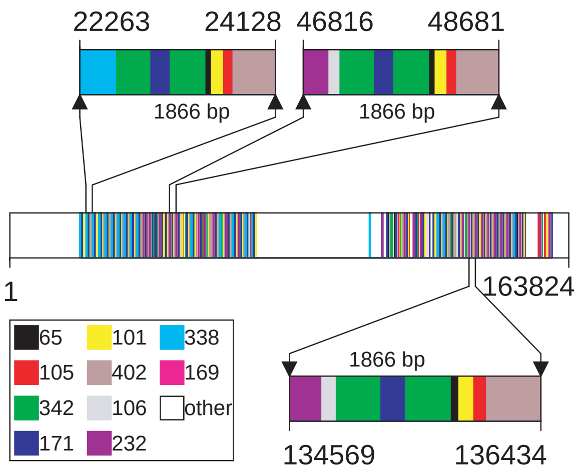

Fig. 4. Colored band for BX284928.3 from chromosome 1. Coloring rule for assignment of particular colors to pronounced subsequence lengths (in bp) is defined in the inset. Positions of nucleotides within BX284928.3 are displayed along the horizontal band (1–163 824). Some intervals displaying color-motif of direct 1866 bp HOR are expanded. Convenient signature of direct HOR: black-yellow-red stripes.

Subsequent to the analysis using the key-string TGTTTC, the analysis using its reverse complement is performed, providing identification and structure of reverse complement HORs. For reverse complement HORs the color-motif is reversed with respect to the color-motif of direct HOR. Accurate visualization of such information allows the user to distinguish direct and reverse complement HORs in a given sequence.

Graphical user interface application for ColorHOR

We have implemented the application associated with ColorHOR for interactive graphical use. The application has been developed using the Python scripting language and the wxPython windows widgets library, and is available at www.hazu.hr/KSA/colorHOR.html. It can be run on Microsoft Windows where wxPython is supported. Technical information for users is provided as Supplementary information. This approach is simple and can be easily incorporated into the analysis of the GenBank database.

In addition to MS Windows binary, we provide a full source code that can be run on any platform with prerequisites satisfied (details in archive) on the project website. Besides MS Windows this application has been tested also on LINUX.

ILLUSTRATION: IMPLEMENTATION OF ColorHOR FOR CHROMOSOME 1

Alpha staircase of chromosome 1

In the first step, we construct the alpha staircase for the GenBank sequence of chromosome 1 using the key-string TGTTTC. For each subsequence length Δ we compute the corresponding frequency N and construct the length–frequency distribution, N versus Δ (Fig. 1). Above the background of noise the most pronounced peaks are associated with 342, 105, 256, 65 and 171 bp subsequences. The noise approximately corresponds to white noise, with superposed very slow decrease of frequency N with increasing subsequence length Δ. Starting from $N \approx 70$ the frequency slowly decreases with Δ, according to a weak exponential dependence $\exp(-k\Delta)$. We note that this noise corresponds to exact matches of the key-string, but at sizeably lower frequencies than the pronounced peaks.

The diagram for cumulative frequency N_c of subsequence length $\Delta = 171$ bp was plotted going along the GenBank sequence of chromosome 1. The computed alpha-staircase (Fig. 2) has a pronounced stair at the position of each alpha satellite region (HOR-type or monomeric type) within the GenBank sequence of chromosome 1. The alpha staircase contains only two pronounced steps, corresponding to the clones (BACs) BX284928.3 (163 725 bp) and BX248407.4 (155 428 bp).

The relationship among monomers within the 1866 bp consensus HOR is investigated by pairwise comparison. The sequence variation among 11 component monomers is 19–46%. In contrast, the average divergence between the 1866 bp HOR copies and consensus HOR is 1.8%, while the average divergence among HOR copies is 2.8%.

ColorHOR display of direct and reverse complement 1866 bp HORs in chromosome 1

After identifying alpha satellite-containing clones in the alpha staircase, we investigate whether these alpha satellites are organized into HORs.

First, we construct the N versus Δ diagram for the clone BX284928.3 (Fig. 3). This clone corresponds to the largest step in the alpha staircase for chromosome 1. The highest peaks are at positions of 342, 171, 65, 105, 101, 402, 106, 232 and 338 bp subsequences. Already in Figure 1, we have identified peaks at these subsequence lengths in the N versus Δ diagram computed for the whole chromosome 1. Additional peaks in Figure 1, at 256 bp and at some shorter subsequence lengths, as well as the background of noise, are not associated with clones BX284928.3 and BX248407.4.

On the basis of pronounced peaks in Figure 3, we define a coloring rule (insets in Figs 4 and 5). This rule assigns a particular color to each subsequence length of pronounced alpha satellite-related peaks in the N versus Δ diagram.

Each of the stripes corresponding to the key-string subsequences along the band displaying the genomic sequence of a clone is colored according to the coloring rule. In this way we construct computationally a colored band. A more detailed insight into HOR structure was achieved by further expanding segments of the colored band.

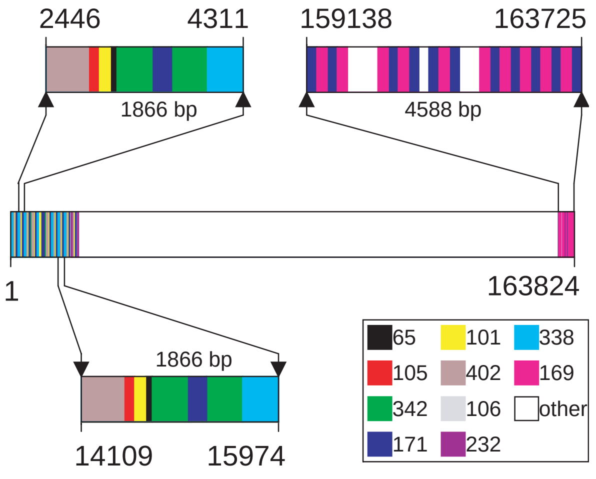

Fig. 5. Colored band for BX284928.3 from chromosome 1 obtained with reverse complement of the key-string TGTTTC. Some intervals displaying color-motif of reverse complement of 1866 bp HOR are expanded. Convenient signature of reverse complement HOR: red-yellow-black stripes, i.e. reverse to signature of direct HORs from Figure 4. With respect to the color-motif of direct HOR, the choice of beginning of reverse complement HOR in expanded view is changed (start: cyan stripe). A different type of reverse complement HOR structure is seen near the end of sequence (expanded, top right): reverse complement of 340 bp dimers (171 bp, 169 bp).

The colored band corresponding to the clone BX284928.3 is displayed in Figure 4 (direct HORs) and Figure 5 (reverse complement HORs).

In the colored band of BX284928.3, corresponding to the key-string TGTTTC, visual inspection reveals a characteristic color-motif of the 1866 bp HOR (magnified segment in Fig. 4, top left). As a signature of this color-motif we choose a pronounced black-yellow-red triplet of stripes.

In the color-motif of some HOR copies the single-base substitutions can lead to fragmentation of a colored stripe into narrower stripes. For example, the 338 bp stripe (cyan) is at some positions fragmented into a 106 bp stripe (gray) and a 232-bp stripe (violet) (magnified segment in Fig. 4, top right), without affecting the global pattern of the color-motif.

In Figure 4 there are two stretches of colored stripes, each of ∼50 kb, with the same characteristic color-motif corresponding to 1866 bp HORs.

In the colored band of BX284928.3 corresponding to the key-string GAAACA (reverse complement of TGTTTC), a visual inspection reveals in the front segment of ∼20 kb a stretch with reverse color-motif (Fig. 5). Each HOR copy in Figure 5 is identical to the reverse complement of the 1866 bp direct HOR from Figure 4 (at the level of 1–3% divergence). This is visually reflected in reversal of the characteristic triplet of stripes: the red-yellow-black triplet of stripes in Figure 5, as the signature of reverse complement HOR structure, is reversed with respect to the black-yellow-red triplet for direct HORs in Figure 4. This provides a simple visual distinction of direct and reverse complement HORs.

Combining Figures 4 and 5 we notice a broad white stripe in the central part of band (roughly between positions 73 000 and 116 000). This ‘hole’ is characterized by LINE/L1, L2 and SINE/Alu GenBank annotation.

In analogy to Figures 4 and 5, we perform the ColorHOR analysis for the GenBank sequence of BX248407.4. We find four distinct stretches of the 1866 bp 11mer HOR sequences, displaying the same color-motif as the clone BX284928.3: two stretches have the color-motif of direct 1866 bp HOR and two of the reverse complement (direct-reverse complement-direct-reverse complement).

Finally, investigating the degree of identity required by the program to declare a match to the key-string, we find that a tandem of alpha monomers is identified as HOR at the level of ∼5% divergence.

KSA segmentation of 1866 bp HORs in chromosome 1 into 11 alpha monomers and novel consensus sequence

Computing the alpha staircase and the ColorHOR diagram in the previous two steps we identified two clones containing HORs, BX284928.3 and BX248407.4. Now we perform a detailed KSA analysis of HORs in BX284928.3 with KSA segmentation of HORs into alpha monomers (Table 1). In supplementary data we display the KSA analysis for BX248407.4 (Table S1). Including all HORs of length 1866 bp (both direct and reverse complement) we derive the 1866 bp consensus HOR (Supplementary data, Table S3). Exactly the same consensus HOR is obtained on replacing the clone BX248407.4 by a more recent version BX248407.26, with KSA analysis displayed in Table S2.

Thus, the 1866 bp HOR is segmented into 11 alpha monomers of lengths 168, 171, 171, 171, 171, 171, 170, 167, 169, 167 and 170 bp.

340 bp HORs in BX284928.3

Near the end of BX284928.3 we found a 4.5 kb stretch characterized by tandem with a color-motif consisting of blue-magenta doublet (magnified segment in Fig. 5, top right). It corresponds to dimers of 340 bp (171 bp, 169 bp). We note that some of the stripes remain blank, as a consequence of point mutations. We identified a stretch of 340 bp dimers. This is probably a front section of a much longer sequence of 340 bp HORs in chromosome 1, extending into a neighboring clone, not yet included in GenBank.

We have also studied the overall mutual sequence divergence of 11mer and dimeric alpha satellite. This divergence is in the range from 23 to 47%.

HOR ANNOTATION FOR WHOLE GenBank SEQUENCE OF HUMAN GENOME AND EXACT CONSENSUS LENGTHS

We apply the ColorHOR algorithm to GenBank data of all human chromosomes. The results of this computational search for HORs with $n > 2$ are displayed in Table 2.

Some of these $n$mers correspond to known $n$mers, previously found by biochemical investigations (Waye *et al*., 1987b; Choo *et al*., 1991; Wevrick and Willard, 1991; Wevrick *et al*., 1992; Warburton and Willard, 1996; Lee *et al*., 1997): 11mers in chromosome 1; 16mers in chromosome 7; 7mers in chromosome 9; 13mers in chromosome 19; and 12mers in chromosome X. Some HORs previously identified by biochemical investigations are not found by computational analysis of GenBank data; this is not surprising having in mind the incompleteness of the centromeric region in GenBank data.

In contrast, we find several new HORs: 13mers in chromosome 4; 11mers in chromosome 8; 4mers in chromosome 9; 18mers in chromosome 10; 12mers in chromosome 11; and 17mers in chromosome 19. These HORs were not reported previously (Warburton and

Table 1. 11mer HOR copies (complete or distorted) identified computationally in the GenBank sequence of BX284928.3 in chromosome 1

HOR	Start	Length	Composition	Divergence(%)
RC	543	1866	11mer	0.91
RC	2409	1866	11mer	0.96
RC	4275	1865	11mer	1.07
RC	6140	1866	11mer	1.02
RC	8006	2037	m01–m07, m07–m11	1.13
RC	10340	1866	11mer	1.13
RC	12206	1866	11mer	1.23
RC	14072	1866	11mer	1.07
RC	15938	1866	11mer	2.09
RC	17804	1866	11mer	2.30
D	20440	1866	11mer	0.59
D	22306	1866	11mer	0.75
D	24172	1866	11mer	1.18
D	26038	1866	11mer	1.39
D	27904	1865	11mer	1.82
D	29769	1866	11mer	1.66
D	31635	1866	11mer	1.39
D	33501	1866	11mer	1.82
D	35367	1866	11mer	1.66
D	37233	1866	11mer	1.02
D	39099	1864	11mer	3.06
D	40963	2714	m01–m08, m04–m11	6.09
D	43897	2715	m01–m08, m04–m11	2.80
D	46859	1866	11mer	3.16
D	48725	1360	m01–m08	5.15
D	50207	1865	11mer	4.66
D	53085	1866	11mer	1.88
D	55219	1866	11mer	4.02
D	57082	1869	ins. CTA	2.78
D	58951	674	m01, m09–m11	1.93
D	59625	1869	ins. CTA	2.30
D	61494	1869	ins. CTA	1.61
D	63363	1869	ins. CTA	2.09
D	65232	1869	ins. CTA	2.25
D	67101	1869	ins. CTA	2.09
D	68970	1868	ins. CTA	1.66
D	70838	1696	m01–m10	1.89
D	115616	1866	11mer	4.77
D	118286	1865	11mer	3.11
D	120419	1869	ins. CTA	2.73
D	122288	1305	m01–m07	8.12
D	123593	1694	m01, m03–m11	1.64
D	125287	1866	11mer	0.91
D	127153	1866	11mer	0.70
D	129019	1865	11mer	1.45
D	130884	1865	11mer	2.41
D	132749	1863	11mer	3.17
D	134612	1866	11mer	2.63
D	136478	1866	11mer	2.84
D	138344	1866	11mer	2.73
D	140210	1866	11mer	2.57
D	142076	1866	11mer	2.84
D	143942	1865	11mer	3.27
D	145807	1865	11mer	3.11
D	147672	1865	11mer	3.06
D	149537	1529	m01-m09	2.55
D	153570	1864	11mer	4.51
D	155434	2712	m01-m09, m05–m11	4.90
D	156962	1184	m05–m11	4.81
D	158146	852	m01–m05	4.81

Notation: D, direct HOR; RC, reverse complement HOR; m01, m02, . . . , m11, monomers in 11mer, defined according to HOR consensus (Table S2) and ins. CTA, insertion of CTA after position 124 in m06. Divergence expressed with respect to consensus HOR was calculated after aligning the corresponding sequences. The string TTCAA is used as a convenient start of alpha monomer units. Monomers in reverse complement HORs are reverse complement with respect to direct HORs.

Willard, 1996; Lee *et al*., 1997). In chromosome 17 we find 14mers; in previous biochemical data (Choo *et al*., 1991) 12mers to 16mers were reported.

For the first time, we determine exact consensus lengths of HORs: 1866 bp in chromosome 1; 2211 bp in chromosome 4; 1868 bp in chromosome 8; 678, 680, 1190 and 1192 bp in chromosome 9; 3058 bp in chromosome 10; 2047 bp in chromosome 11; 2379 bp in chromosome 17; 2896 and 2214 bp in chromosome 19 and 2057 bp in chromosome X.

DISCUSSION

One should take into account several issues concerning genome assembly and known higher-order alpha satellites. For example, HOR-containing BACs BX284928.3 and BX248407.4 in chromosome 1 were of working draft quality, composed of unordered pieces, not fully assembled and not included in the contiguous map. As such, they do not connect to any other contiguous chromosome 1 scaffold and are surrounded by genome gaps on both sides. It is even questionable whether these gaps are distinct, or whether they overlap each other and are in part duplicates of the same region of genome. We investigated this problem by identifying complete overlaps (100% convergence) (direct, reverse, complement or reverse complement) of intervals >1 kb. In this way we identified seven overlap intervals (Supplementary data, Table S4). These intervals are interspersed and mapped in an irregularly intertwined way. For example, the interval from position 37992 to 42322 in BX284928.3 is mapped into reverse complement of interval from 82472 to 86802 in BX248407.4. The longest overlap interval is of 23 276 bp. The total length of seven overlap intervals of these two BACs is 66 773 bp.

Furthermore, it should be noted that the sequence BX248407.4 was released in 2003 and consists of 11 unordered pieces. We performed an additional analysis, using the completed sequence BX248407.26, released in 2004. We find that HORs which are reverse complemented in BX248407.4 turn into direct HORs in BX248407.26.

Using the ColorHOR program we obtain a simple graphical display of direct HORs, corresponding to the characteristic color-motif, and of reverse complement HORs, corresponding to the reverse color-motif. We note that inversions in alpha satellite DNA have been previously found in chromosome 7 (Wevrick *et al*., 1992), where two

Table 2. HOR annotation of the GenBank data for human genome using ColorHOR algorithm

Chromosome	Clone	ColorHOR results		Previous biochemical data
		n mer	Consensus length (bp)	[HOR length (*n*mer)]
1	BX248407.7[d]	11mer	1866	1.9 kb (11mer)[a,b]
	BX284928.3[d]	11mer	1866	
4	AC027271.7	13mer	2211	3.4 kb (20mer)[a], 2.6 kb (15mer)[a], 3.2 kb (19mer)[b], 1.2 kb (7mer)[b]
7	AC017075.8	16mer	2734	2.7 kb (16mer)[a,b], 1.02 kb (6mer)[c]
	AC133532.1[d]	16mer	2734	
8	AC118650.5	11mer	1868	2.5 kb (15mer)[a,b]
	AC144576.2	11mer	1868	
9	BX088702.4[d]	4mer	678	2.7 kb (16mer)[a,b], 1.2 kb (7mer)[b]
		4mer	680	
		7mer	1190	
		7mer	1192	
10	BX322613.6	18mer	3058	1.35 kb (8mer)[a,b]
11	AC126345.11	12mer	2047	0.85 kb (5mer)[a,b]
17	AC131274.9	14mer	2379	2.05–2.74 kb (12mer to 16mer)[c]
	AC145160.1[d]	14mer	2379	
19	AC073541.4	17mer	2896	
	AC136499.2[n]	13mer	2214	2.25 kb (13mer)[a,b]
X	AL591645.35	12mer	2057	2.0 kb (12mer)[a,b]
	BX537339.3	12mer	2057	

Last column: Previously reported typical HOR lengths in chromosomes 1, 4, 7, 8, 9, 10, 11, 17, 19 and X from compilations of [a]Lee *et al*. (1997); [b]Warburton and Willard (1996) and [c]Choo *et al*. (1991); [d]in draft stage; [n]not placed.

of four examined clones contained large inversions. This was unexpected, as all thoroughly analyzed alpha satellite clones previously isolated have been tandem and unidirectional.

We performed a sequence divergence study taking into account that HORs in either forward or reverse orientations are not distinct species, and that, given the lack of assembly validation, one cannot be sure of orientation. Therefore, the only divergence of consequence is overall sequence divergence comparing all 11mers in both BACs in chromosome 1. The calculated divergence is somewhat higher than the previously determined sequence variation of 16–36% among monomers of pSD1-1 (Waye *et al*., 1987). In contrast, we compared our 1866 bp consensus HOR to the previously determined 11mer in chromosome 1, pSD1-1 (Willard and Waye, 1987b). Aligning these sequences we found that they diverge by 9%, showing a sizeable degree of similarity. Furthermore, we calculated the divergence of 11mer and dimeric alpha satellites, showing that dimeric alpha satellite in BX284928.3 is a distinct type of HOR from 11mer.

Repeat units of *n* bp arranged in a head-to-tail fashion have an arbitrary beginning or end that can be described as starting at any of the *n* registers. The historical use of location of restriction enzyme sites has resulted in different published starts of homologous repeat units (Warburton *et al*., 1993). In this sense, a choice of key-string in the KSA algorithm could be compared to a choice of a particular ‘computer enzyme’.

Our graphical computational method is very efficient and simple from a computational viewpoint, with powerful graphical user interface applications. It provides a fast scan of HORs in a whole chromosome sequence, taking ∼1 min computing time per chromosome using PC Pentium IV. Our HOR-searching method does not involve any numerical parameter, the only input being the dominant key-string, which is deduced as a robust string from known alpha satellite consensus.

Let us comment on a comparison with the well-known computational tool Tandem repeats finder (Benson, 1999). Our method is very fast and selective in screening for HORs in very long sequences, of hundreds of Mb. It also provides identification of individual HORs outside of tandem, that are not identified by Tandem repeats finder. Furthermore, the most recent version of Tandem repeats finder (ver. 3.21) could identify HORs up to the length of 2000 bp, with a suitable choice of parameters, while our method has no such limitation and can identify HORs of any length. We also point out that the KSA analysis provides a full list and structure of insertions, deletions and point mutations within HORs, and identifies HORs in the presence of more pronounced insertions and/or deletions. For example, in BX284928.3 our method identifies 60 HORs (complete or incomplete), while the Tandem repeats finder identifies 51 HORs.

We note that the ColorHOR program can be used to analyze other satellite sequences as well, as for example monomeric alpha satellites. This program could also be used for the identification of interspersed repeats by using proper robust strings.

This paper shows that it is possible to construct a simple graphical representation displaying the HOR structure within an arbitrarily large genomic sequence, like DNA sequence of a whole human chromosome. Full sequencing of the human genome (sizeable segments from the centromeric region are still missing or in draft stage) will provide the opportunity for computational identification and analysis of all HORs, leading to complete HOR-annotation of human genome.

Finally, we note that the HOR classification of genomic sequences has not yet been incorporated into the GenBank database. We suggest that the HOR assignment obtained by ColorHOR graphical user interface application be included into GenBank class of repetitions. We also suggest the ColorHOR scanning of all clones in the draft stage, not yet included into GenBank.

ACKNOWLEDGEMENT

We thank H. F. Willard for valuable comments.

REFERENCES

Alexandrov,I.A., Medvedev,L.I., Mashkova,T.D., Kisselev,L.L., Romanova,L.Y. and Yurov,Y.B. (1993) Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization. *Nucleic Acids Res.*, **21**, 2209–2215.

Baldi,P. and Baisnee,P.F. (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics*, **16**, 865–889.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Benson,G. and Waterman,M.S. (1994) A method for fast database search for all k-nucleotide repeats. *Nucleic Acids Res.*, **22**, 4828–4836.

Blanchard,M.K., Chiapello,H. and Coward,E. (2000) Detecting localized repeats in genomic sequences: a new strategy and its applications to *Bacillus subtilis* and *Arabidopsis thaliana* sequences. *Comput. Chem.*, **24**, 57–70.

Borštnik,B., Pumpernik,D., Lukman,D., Ugarković,D. and Plohl,M. (1994) Tandemly repeated pentanucleotides in DNA sequences of eukaryotes. *Nucleic Acids Res.*, **22**, 3412–3417.

Castello,A.T., Martins,W. and Gao,G.R. (2002) TROLL–Tandem Repeat Occurrence locator. *Bioinformatics*, **18**, 634–636.

Choo,K.H., Vissel,B., Nagy,A., Earle,E. and Kalitsis,P. (1991) A survey of the genomic distribution of alpha satellite on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Res.*, **19**, 1179–1182.

Goodstadt,L. and Ponting,C.P. (2001) CHROMA: consensus-based coloring of multiple alignments for publication. *Bioinformatics*, **17**, 845–846.

Hauth,A.M. (2002) Identification of tandem repeats: simple and complex pattern structures in DNA sequences. Dissertation, University of Wisconsin-Madison.

Hauth,A.M. and Joseph,D.A. (2002) Beyond tandem repeats: complex pattern structures and distant regions of similarity. *Bioinformatics*, **S18**, 31–37.

Lee,C., Wevrick,R., Fisher,R.B., Ferguson-Smith,M.A. and Lin,C.C. (1997) Human centromeric DNAs. *Hum. Genet.*, **100**, 291–304.

Maio,J.J. (1971) DNA strand reassociation and polyribonucleotide binding in the African green monkey, *Cercopithecus aethiops*. *J. Mol. Biol.*, **56**, 579–595.

Manuelidis,L. and Wu,J.C. (1978) Homology between human and simian repeated DNA. *Nature*, **276**, 92–94.

Milosavljevic,A. and Jurka,J. (1993) Discovering simple DNA sequences by the algorithmic significance method. *Comput. Appl. Biosci.*, **9**, 407–411.

Puente,A. de la, Velasco,E., Perez Jurado,L.A., Hernandez Chico,C., Rijke,F.M. van de, Scherer,S.W., Raap,A.K. and Cruces,J. (1998) Analysis of the monomeric alphoid sequences in the pericentromeric region of human chromosome 7. *Cytogenet. Cell Genet.*, **83**, 176–181.

Romanova,L.Y., Deriagin,G.V., Mashkova,T.D., Tumeneva,I.G., Mushegian,A.R., Kisselev,L.L. and Alexandrov,I.A. (1996) Evidence for selection in evolution of alpha satellite DNA: the central role of CENP-B/pJα binding region. *J. Mol. Biol.*, **261**, 334–340.

Rosandić,M., Paar,V. and Basar,I. (2003) Key-string segmentation algorithm and higher-order repeat 16mer (54 copies) in human alpha satellite DNA in chromosome 7. *J. Theor. Biol.*, **221**, 29–37.

Schueler,M.G., Higgins,A.W., Rudd,M.K., Gustashaw,K. and Willard,H.F. (2001) Genomic and genetic definition of a functional human centromere. *Science*, **294**, 109–115.

Taylor,W.R. (1986) The classification of amino acid conservation. *J. Theor. Biol.*, **119**, 205–218.

Warburton,P.E. and Willard,H.F. (1996) Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes. In Jackson,M., Strachan,T. and Dover,G. (eds.), *Human Genome Evolution*. BIOS, Oxford, pp. 121–145.

Warburton,P.E., Waye,J.S. and Willard,H.F. (1993) Nonrandom localization of recombination events in human alpha satellite repeat unit variants: implications for higher-order structural characteristics within centromeric heterochromatin. *Mol. Cell. Biol.*, **13**, 6520–6529.

Waye,J.S. and Willard,H.F. (1985) Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome. *Nucleic Acids Res.*, **13**, 2731–2743.

Waye,J.S. and Willard,H.F. (1987) Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res.*, **15**, 7549–7569.

Waye,J.S., England,S.B. and Willard,H.F. (1987a) Genomic organization of alpha satellite DNA on human chromosome 7: evidence for two distinct alphoid domains on a single chromosome. *Mol. Cell Biol.*, **7**, 349–356.

Waye,J.S., Durfy,S.J., Pinkel,D., Kenwrick,S., Patterson,M., Davies,K.E. and Willard,H.F. (1987b) Chromosome-specific alpha satellite DNA from human chromosome 1: hierarchical structure and genomic organization of a polymorphic domain spanning several hundred kilobase pairs of centromeric DNA. *Genomics*, **1**, 43–51.

Wevrick,R. and Willard,H.F. (1991) Physical map of the centromeric region of human chromosome 7: relationship between two distinct alpha satellite arrays. *Nucleic Acids Res.*, **19**, 2295–2301.

Wevrick,R., Willard,V.P. and Willard,H.F. (1992) Structure of DNA near long tandem arrays of alpha satellite DNA at the centromere of human chromosome 7. *Genomics*, **14**, 912–923.

Willard,H.F. (1985) Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.*, **37**, 524–532.

Willard,H.F. and Waye,J.S. (1987a) Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.*, **3**, 192–198.

Willard,H.F. and Waye,J.S. (1987b) Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J. Mol. Evol.*, **25**, 207–214.